VocalPrint: Exploring A Resilient and Secure Voice Authentication via mmWave Biometric Interrogation

Huining Li[1], Chenhan Xu[1], Aditya Singh Rathore[1], Zhengxiong Li[1], Hanbin Zhang[1], Chen Song[2], Kun Wang[3], Lu Su[1], Feng Lin[4], Kui Ren[4], Wenyao Xu [1]
[1]University at Buffalo, the State University of New York, Buffalo, New York, USA
[2]San Diego State University, San Diego, California, USA
[3]University of California, Los Angeles, California, USA
[4]Zhejiang University, Zhejiang, China
{huiningl,chenhanx,asrathor,zhengxio,hanbinzh,lusu,wenyaoxu}@buffalo.edu,
csong@sdsu.edu,wangk@ucla.edu,{flin,kuiren}@zju.edu.cn

ABSTRACT

With the continuing growth of voice-controlled devices, voice metrics have been widely used for user identification. However, voice biometrics is vulnerable to replay attacks and ambient noise. We identify that the fundamental vulnerability in voice biometrics is rooted in its indirect sensing modality (*e.g.*, microphone). In this paper, we present *VocalPrint*, a resilient mmWave interrogation system which directly captures and analyzes the vocal vibrations for user authentication. Specifically, *VocalPrint* exploits the unique disturbance of the skin-reflect radio frequency (RF) signals around the near-throat region of the user, caused by the vocal vibrations during communication. The complex ambient noise is isolated from the RF signal using a novel resilience-aware clutter suppression approach for preserving fine-grained vocal biometric properties. Afterward, we extract the text-independent vocal tract and vocal source features and input them to an ensemble classifier for user authentication. *VocalPrint* is practical as it leverages a low-cost, portable, and energy-efficient hardware allowing effortless transition to a smartphone while having sufficient usability as typical voice authentication systems due to its non-contact nature. Our experimental results from 41 participants with different interrogation distances, orientations, and body motions show that *VocalPrint* can achieve over 96% authentication accuracy even under unfavorable conditions. We demonstrate the resilience of our system against complex noise interference and spoof attacks of various threat levels.

CCS CONCEPTS

• **Human-centered computing** → **Ubiquitous and mobile computing**; • **Security and privacy** → **Biometrics**.

KEYWORDS

Voice authentication; mmWave sensing

SenSys '20, November 16–19, 2020, Virtual Event, Japan

ACM ISBN 978-1-4503-7590-0/20/11…$15.00
https://doi.org/10.1145/3384419.3430779

ACM Reference Format:
Huining Li[1], Chenhan Xu[1], Aditya Singh Rathore[1], Zhengxiong Li[1], Hanbin Zhang[1], Chen Song[2], Kun Wang[3], Lu Su[1], Feng Lin[4], Kui Ren[4], Wenyao Xu [1]. 2020. VocalPrint: Exploring A Resilient and Secure Voice Authentication via mmWave Biometric Interrogation . In *The 18th ACM Conference on Embedded Networked Sensor Systems (SenSys '20), November 16–19, 2020, Virtual Event, Japan.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3384419.3430779

1 INTRODUCTION

Due to the growth of voice-controlled devices and services, the use of vocal-based biometrics for user authentication has surged [18, 37]. Voiceprint is a strong physiological and behavioral combined biometrics, considered to be just as biologically unique in individuals as a fingerprint [26]. There is a sizable literature on user identification by analyzing voice, including speech [8] and non-speech [61] vocal data. Commodity voice-controlled devices, such as the Amazon Echo and Google Home, have integrated speaker identification functions to secure the user information [29].

However, there are several major security limitations for adopting voice biometric technologies in real-world applications [63]. For example, fraudsters may eavesdrop the legitimate user's speech samples or utilize a variety of artificial intelligence technologies to generate synthetic voice data [73], and then launch a "replay attack" against voice-based authentication systems. How to defend against the playback attack has a long and rich history, and is a core research problem in biometric security [75]. Researchers have studied sets of software-based solutions based on liveness distinction between human and loudspeakers, including challenge-response protocols [5], ultrasonic reflections of mouth motion [76], time-difference-of-arrival (TDoA) of phenome sounds to two microphones [77], sound field difference [71], etc. Although these approaches could alleviate the security risk under some circumstances, they need user's active cooperation and also assume that replaying cannot generate identical sound waves. We discover that the fundamental vulnerability in voice biometrics is rooted in its **indirect sensing modality**. Currently, voice biometric systems mainly employ a microphone sensor. When the user speaks, the vocal folds vibrate. The generated sound propagates in the air media, and the air vibration is captured by a microphone. This kind of indirect voice sensing modality through a media creates an inevitable attack surface in

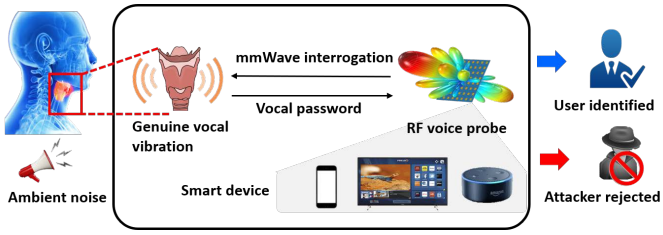

Figure 1: A mmWave biometric interrogation system leveraging the vocal vibrations for accurate user authentication.

the physical world (*e.g.*, replay attacks using a high-definition loudspeaker or bionic loudspeaker arrays), which is hardly addressed by software-based approaches. Moreover, indirect voice sensing is prone to interference from ambient noises which decrease the usability of voice biometric systems, as it leads to false positives during authentication and is insensitive to minute alteration in fake voice input during replay attacks.

It is a fact that voiced sound is determined by vocal fold vibration, which is the root of voiceprint uniqueness [26]. On the basis of this argument, we propose that the most secure and attack-resistant voice sensing approach to user identification is to **directly** acquire and analyze the user's vocal fold vibration. Radio-frequency (RF) signals, such as millimeter wave (mmWave), shows immense potentials in sensing micron-level skin displacement [57, 68] due to their directional beamforming and skin-reflectance properties. A recent study demonstrated the feasibility of acquiring the vocal vibrations that occur at the range of 2-3mm via mmWave radar [33]. Motivated by these works, a **non-contact** and **direct** biometric mmWave interrogation system can be developed to capture the unique vocal vibrations for secure user authentication.

To realize our system, we need to address the following challenges: (1) How to suppress the complex noise clutters arising from static and dynamic objects in the environment and motion artifact for preserving fine-grained voice biometric properties in received mmWave response? (2) How to extract and identify the intrinsic features that can perfectly capture the vocal tract and vocal source information to maximize the system performance? (3) How to validate the resilience of our system against spoof attacks?

To this end, we present our system, *VocalPrint*, to facilitate a resilient mmWave interrogation system for secure and non-contact voice authentication, illustrated in Figure 1. We leverage a low-cost, portable, and high-resolution 77GHz Frequency Modulated Continuous Wave (FMCW) radar to identify the user from the dynamic environment and non-invasively sense the minute vocal vibrations. The displacement in the vocal vibrations is inferred from the phase shift of the peak corresponding to the human target in the intermediate frequency (IF) signals. To help reserve fine-grained voice biometric properties in the RF voice signals, we develop a resilient-aware assembled clutter suppression scheme to isolate random motion artifact and ambient noise clutter. Once the precise vocal vibration signals are obtained, we extract text-independent vocal source and vocal tract features, respectively, which closely relate to the human speech articulation. Finally, these text-independent biometric descriptors are fed into a fine-tuned feature selection module and an ensemble classifier for user authentication. To intensively evaluate our system, we recruit 41 volunteers with results showing that *VocalPrint* can enable a reliable authentication with over 96% accuracy. Furthermore, we validate the resilient security of *VocalPrint* against ambient interference (*e.g.*, acoustic noise, dynamic environment, human obstruction) and spoofing attacks (*e.g.*, counterfeit, mimicry, signal-based) to show its significant potential as an enhancement to voice authentication in real-world applications.

The contribution of our work has three-fold:

- We perform the first study to identify that the fundamental vulnerability in voice biometrics is rooted in its indirect sensing modality. We also explore a direct mmWave sensing approach to acquire and analyze the user's vocal fold vibration in a secure and attack-resistant manner.
- We develop *VocalPrint*, an end-to-end biometric system to facilitate resilient security of voice authentication. We first design a novel resilience-aware clutter suppression model to obtain precise vocal vibration data that reserves fine-grained biometric properties, and then extract intrinsic features that depict vocal source and vocal tract information for user identification.
- We demonstrate the effectiveness and robustness of *VocalPrint* through extensive experiments with results showing superior authentication accuracy even under unfavorable conditions. We conduct comprehensive studies to validate the resilience of *VocalPrint* against complex noise interference and spoof attacks of various threat levels.

2 THEORY AND PRELIMINARIES

2.1 Voice Biometrics Rationale

Voice can be regarded as physiological and behavioral combined biometrics, which contains unique and permanent information of individuals [26]. Specifically, voice permanence is derived from the fixed physical shape of individual's lung, vocal cords, and vocal tract. Voice uniqueness stems from the precise and coordinated vibration of the vocal cords and vocal tract [17, 55]. When a person speaks, the air flow is first expelled from the lungs and then traverses through the vocal cords. The vocal cords with the glottis constrict to block the air flow and the resulting vibrations in air produce voiced signals. In contrast, when the vocal cords with the glottis dilate, the air flow is allowed to pass through without heavy vibrations, thereby generating unvoiced signals. Afterward, both voiced and unvoiced signals are resonated and reshaped by the vocal tract consisting of multiple articulatory organs (*e.g.*, epiglottis, corniculate cartilage, cuneiform cartilage, shown in Figure 2). The movement of articulatory organs forms a path with specific geometrical shapes (*i.e.*, articulatory gesture) for the air flow [27], which manipulates the amplitude and frequency of vocal vibrations. Although different people may share the same type of articulatory gesture when pronouncing the same phoneme, the movement speed and intensity vary from person to person and contain distinctive information. Moreover, the larynx modulates the tension on vocal cords to produce fine-tuned vocal vibrations, which further adds the uniqueness to an individual voice. Therefore, this uniqueness of the human voice is intrinsically sourced in vocal vibrations.

2.2 A Preliminary Study

There is a significant growing interest in human sensing applications using RF sensing [30, 36, 65]. Specifically, WaveEar [69] is one recent representative work on investigating speech recognition

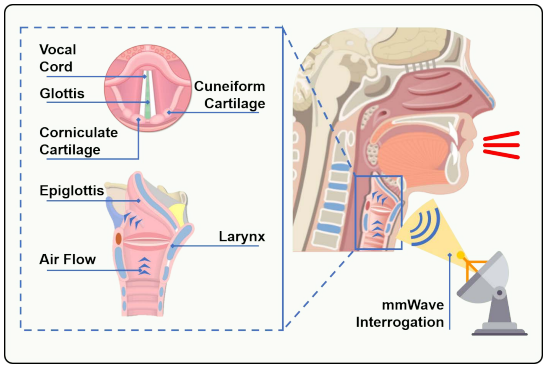

Figure 2: The vocal vibrations result from dynamics of articulatory organs and can be sensed by the mmWave radar from near-throat region.

using mmWave technologies. To examine the feasibility of acquiring vocal biometric features in mmWave sensing, we conduct a preliminary study using a mmWave-band FMCW probe.

Preliminary Data Collection. In the preliminary experiment, we leverage a beamforming mmWave probe to sense the subject's vocal vibration and collect received mmWave signals. Specifically, two subjects are asked to sit in the same position and pronounce the sentence, "After class, he went home". For the ease of analysis, we align the mmWave probe in the direction of the subject's throat. The distance between the subject and the probe is 20cm.

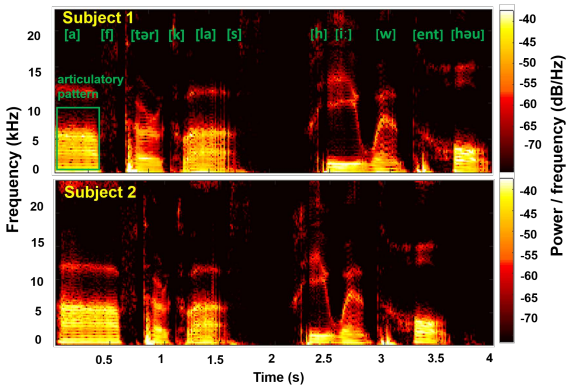

Figure 3: Spectrogram of reconstructed voice.

Auditory Analysis in mmWave Sensing. Speech recognition [60] and speaker identification [54] are both voice-based applications. However, underlying mechanisms and associated technologies are distinct. Speech recognition utilizes the temporal cues and envelopes in voice data and it is critical to capture and parse coarse-grained (*e.g.*, hundreds of milliseconds to seconds) articulatory features (*e.g.*, up/down/back/forth movement). Speaker identification exploits spectral information in voice data. For example, fine-grained (*e.g.*, tens of milliseconds) spectral envelopes contain the resonance properties of vocal tracts and timber, which are the pivotal features in speaker identification. We adopted the analytical scheme in WaveEar [69] to reconstruct voice signals of both speakers. As shown in Figure 3, both reconstructed voices have a similar spectrum and envelope (with a segment of 100ms). The voice data can be successfully processed by the commodity speech recognition software kit [50]. However, as shown in Figure 4, the short-term (10ms) spectral envelopes in both reconstructed voices have a low resolution, and spectral poles in both spectrums are nearly the same. Biometric traits are lost in the mmWave-reconstructed voice data.

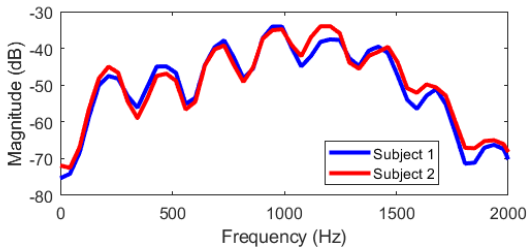

Figure 4: Short-term spectral envelop.

Summary: A new analytical scheme in processing mmWave signals is investigated for speaker identification. Particularly, short-term proprieties in vocal patterns need to be reserved and augmented. In the following sections, we will present (A) high-definition mmWave interrogation, (B) a resilience-aware clutter removal scheme using a novel assembled model, and (C) robust feature extraction and matching methods.

3 VOCALPRINT SYSTEM OVERVIEW

In this paper, we introduce *VocalPrint*, a resilience-aware mmWave biometric interrogation system. The end-to-end system overview is shown in Figure 5.

***VocalPrint* Hardware**: A high-resolution mmWave probe is leveraged to accurately sense the vocal vibrations in a non-contact manner. Specifically, the probe first transmits a frequency modulated continuous wave towards the throat of the user and then receives the skin-reflect response signal which comprises sufficient information of the vocal vibrations when the user is speaking. The received signal is transmitted to a resilience-aware clutter suppression model for isolating the noise from the surrounding environment and even the dynamic obstruction caused by multiple human subject interference.

***VocalPrint* Software**: Once the precise vocal vibration data is acquired, *VocalPrint* extracts optimal biometric features that depict vocal source and vocal tract information. After that, the vocal biometric descriptors are input to a fine-tuned authentication model that consists of a feature selection module and an ensemble classifier for identifying the legitimate user against imposters.

4 MMWAVE INTERROGATION OF VOCAL VIBRATIONS

4.1 mmWave Probe Design and Integration

The Continuous Wave (CW) is increasingly used in sensing various vital signs, such as breathing and heartbeat, due to its ability to capture near-field motion and displacement [34]. However, CW is not accurate in range measurement because it lacks the timing mark (the frequency is fixed). Besides, CW cannot differentiate between two or more reflecting objects because the reflected signals and clutters are all mixed up in both the time and frequency domains. Therefore, we conclude that CW is not capable of authenticating a person at the non-pre-known position in a complex environment. In *VocalPrint*, we leverage FMCW, which can detect both accurate range and minute displacement. Moreover, FMCW enables a low-frequency received signal processing by the mixed IF signal, which

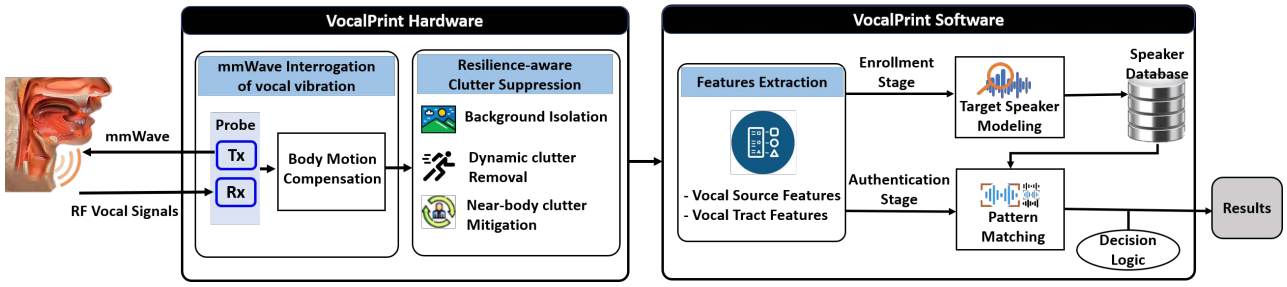

Figure 5: The overview of *VocalPrint* mainly consisting of a mmWave interrogation module to sense the vocal vibrations, a resilience-aware clutter suppression module to remove the complex noise, and an authentication module to identify the legitimate user against imposters.

considerably reduces the loading of designing and realizing the circuit. In the next part, we give the formal description of the continuous vocal vibration interrogation utilizing the FMCW mmWave. Based on the interrogation theory, we give more discussion about the mmWave probe parameters in Section 7.1.

4.2 Continuous Vocal Vibration Interrogation

To enable the continuous vocal vibration interrogation, FMCW modulates a saw-tooth baseband (used as timing mark) to the high-frequency mmWave carrier. Specifically, the periodic chirp signal $T(t)$ transmitted to the speaking person's throat is defined as:

$$T(t) = \exp\left[j\left(2\pi f_0 t + \int_0^t 2\pi\rho t\, dt\right)\right], \tag{1}$$

where $0 < t < T_r$, f_0 is the carrier frequency, T_r is one chirp cycle, B is the bandwidth of one chirp, and $\rho = B/T_r$ is the chirp rate. Assume that the distance between the radar and human throat is $X(t) = X_0 + d(t)$, where X_0 is the original distance, $d(t)$ represents the minute skin displacement caused by vocal vibration. With the round trip delay t_d lagged behind the transmitted chirp signal, the received signal consists of the clutter components $R_{\text{clutter}}(t)$ and the vocal component $R(t)$ carrying the vocal vibration, which is:

$$R(t) = \Gamma \exp[j(2\pi f_0(t - t_d) + \int_0^{t-t_d} 2\pi\rho t\, dt)], \tag{2}$$

where $t_d < t < T_r + t_d$, Γ denotes the amplitude normalized to the transmitted chirp signal, and $t_d = \frac{2[X_0+d(t)]}{c}$. The clutter suppression is studied further in Section 5.

For every chirp, the valid time period for mixing is (t_d, T_r), and thereby the IF signal for a chirp after mixing can be obtained as:

$$H(t) = T(t) \times R^*(t) \approx \Gamma \exp[j(2\pi\rho t_d t + 2\pi f_0 t_d)], \tag{3}$$

where $*$ represents a conjugate transpose operation, $\times$ is the mixer, the mathematical term related to t_d^2 is left out due to $t_d^2 << t_d t$, $t_d < t < T_r$. From Eq. (3), we can see that the mixed IF signal is directly related to the skin displacement caused by vocal vibration $d(t)$. Since $d(t)$ is very small during one chirp, we track the IF signal across a sequence of M chirps. Substituting $t_d = \frac{2[X_0+d(t)]}{c}$, the IF signal for the m-th chirp period can be formulated as:

$$\begin{aligned} H(mT_r + t) = \Gamma \exp\{j[&\frac{4\pi\rho X_0}{c} t + \frac{4\pi f_0 X_0}{c} \\ &+ (\frac{4\pi\rho t}{c} + \frac{4\pi f_0}{c}) d(mT_r)]\}, \end{aligned} \tag{4}$$

where c denotes the light speed. Because of $\rho t << f_0$ ($t \in (t_d, T_r)$) in typical FMCW radars, $\frac{4\pi\rho t}{c}$ can be neglected. Then, $H(mT_r + t)$ can be obtained as:

$$\begin{aligned} &H(mT_r + t) = \Gamma \exp[j(w_H t + \psi_m)], \\ &w_H = \frac{4\pi\rho X_0}{c},\ \psi_m = \frac{4\pi f_0 X_0 + 4\pi f_0 d(mT_r)}{c}. \end{aligned} \tag{5}$$

Therefore, the vibration displacement during the m-th chirp period $d(mT_r)$ can be calculated as:

$$d(mT_r) = \frac{c}{4\pi f_0}\Delta\psi_m, \tag{6}$$

where $\Delta\psi_m$ can be achieved by conducting Fast Fourier Transform (FFT) on the IF signals for a sequence of M chirps.

4.3 Body Motion Compensation

Random body motion from users is the main barrier in achieving precise vocal vibration data. When the body motion amplitude is more than half of the range profile resolution ΔRES ($\Delta RES = \frac{c}{2B}$), the range bins will be misaligned, and hardly calculate accurate $\Delta\psi_m$. A conventional solution is to use a digital filter and compensate for the body motion accordingly, but it may cause cumulative errors over time. To address this issue, we develop a fine-grained range bin alignment solution.

We define $S_m(l)$ as the m-th acquired range profile, where $m = 0, ..., M-1$, $l = 0, ..., L-1$, M is the number of the acquired range profiles (*i.e.,* the number of chirps), and L is the number of range bins. $\widetilde{S}_m(l)$ is denoted as the aligned range profile, χ_m is denoted as the range shift added to the $S_m(l)$, and $S_m(l - \chi_m)$ represents the shifted range profile of $S_m(l)$. We also define the reference range profile that exploits the knowledge of the previously aligned range profiles, formulated as:

$$Q_{m+1}(l) = \frac{m}{m+1} Q_m(l) + \frac{1}{m+1}|\widetilde{S}_m(l)|. \tag{7}$$

In the beginning, we consider $Q_m(0) = \widetilde{S}_m(0) = S_m(0)$.

To align a sequence of M range profiles $S_{m+1}(l)$, we formulate the envelope correlation function between the shifted range profile and its corresponding reference range profile, denoted as:

$$\Pi(\chi_{m+1}) = \sum_{l=0}^{L-1} |Q_{m+1}(l)| \cdot |S_{m+1}(l - \chi_{m+1})|. \tag{8}$$

The maximum value of $\Pi(\chi_{m+1})$ indicates the optimum alignment between the $(m+1)$-th shifted range profile $S_{m+1}(l - \chi_{m+1})$ and the $(m+1)$-th reference range profile $Q_{m+1}(l)$. Therefore, we first calculate the integer value of χ_{m+1} ($\chi_{m+1} \in [0, 1, ..., L-1]$) to

maximize the value of $\Pi(\chi_{m+1})$, and denote this value as χ^0_{m+1}. Then, we deploy the Nelder-Mead algorithm [28] to explore optimum range shift χ^{opt}_{m+1} for achieving local maximum, and χ^0_{m+1} is taken as an initial guess for the exploration. Finally, the $(m+1)$-th aligned range profile can be obtained as:

$$\begin{aligned}\widetilde{S}_{m+1}(l) &= S_{m+1}(l - \chi^{opt}_{m+1}) \\ &= \text{FFT}\{\exp(j(2\pi \frac{\chi_{m+1}}{L} \triangle)\text{IFFT}\{S_{m+1}(l)\}\},\end{aligned} \tag{9}$$

where $\triangle$ is the vector $[0, 1, ..., L-1]^\top$. After finishing this process, we align the next range profile.

Preliminary Results: In our preliminary work, we collected the data of the body motion artifacts from a mmWave hardware platform and simulated the proposed method. The subject is asked to sit in the direction of the mmWave probe and randomly wobble his upper body when speaking. Other experiment settings are the same as we mentioned in Section 2.2. Figure 6 shows the range-Doppler-matrix (RDM) before (a) and after (b) body motion removal. RDM is the result of the frequency domain analysis among multiple range profiles, which can illustrate noises and signal-of-target. It indicates that our proposed method can eliminate interference from random body motion. Note that in this simulated study, the clutter on the background still exists. Next, we will introduce the resilience-aware clutter suppression approaches to enhance voice features in RF voice data.

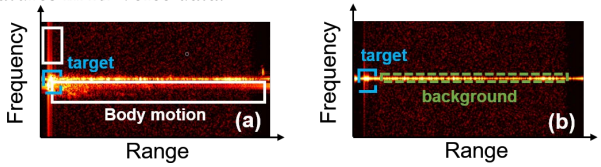

Figure 6: The RDM before (a) and after (b) the body motion compensation.

5 RESILIENCE-AWARE CLUTTER SUPPRESSION

Undesirable backscatters caused by static/dynamic surrounding objects become barriers in precision voice acquisition that can disturb short-term spectral properties. Therefore, we investigate a resilience-aware clutter removal scheme using a novel assembled model to preserve voice biometric features in RF streaming signals.

5.1 Background Clutter Isolation

The background clutter in reflected mmWave signals is more complicated than that in conventional acoustic signals. Specifically, outdoor (*e.g.,* snow and rain) and indoor (*e.g.,* furniture and computer) background are both able to reflect the high-frequency mmWave whose coverage range reaches 100 m [2]. With different reflection rate to mmWave and the multipath interference, these objects are not able to be isolated by simply applying a threshold or training a classifier (which also brings much more overhead). Therefore, we regard the background clutter in *VocalPrint* as the accumulation of the reflected signal by many small parts of the background, such as table legs, chairs, and monitors. The amplitude and phase of these reflections are random and its spectrum envelope a obeys to the following Weibull distribution [46] about the clutter:

$$p(a) = \frac{na^{n-1}}{\mu^n} \exp\left[-\left(\frac{a}{\mu}\right)^n\right], \tag{10}$$

where n and μ are the shape and scale parameters of the distribution, respectively.

To isolate the background clutter, we arrange the range profiles $\widetilde{S_m}, (m = 1, 2, \cdots, M)$ that are obtained in Eq. (9) to a matrix (M rows $\times$ L columns) and perform the second FFT chirp-wise (slow-time FFT) to get the range-Doppler-matrix (RDM). Then target searching (isolation) is performed on the log-normalized RDM to isolate the background clutter. Here, we define a resilient matrix u as $u = \text{C}(\mathcal{R})$, where $\mathcal{R}$ represents the RDM. Because of the complexity of the background clutter, the resilient function C is an $M \times L$ matrix of functions in which each element can be formulated as:

$$c_{ij} = \frac{\ln|\mathcal{R}_{ij}| - \hat{\mathbb{E}}(a)}{\text{std}_a}, \tag{11}$$

where $\text{std}_a = \frac{\pi}{n\sqrt{6}}$. $\hat{\mathbb{E}}(a)$ is the unbiased estimation of the $\mathbb{E}(a)$ [70], which can be calculated by:

$$\hat{\mathbb{E}}(a) = \frac{1}{10}\Big(\sum_{i-12,j-3}^{i+12,j+3} |\mathcal{R}_{ij}| - \sum_{i-10,j-2}^{i+10,j+2} |\mathcal{R}_{ij}|\Big), \tag{12}$$

where 12, 3, 10, 2 are empirical value. Finally, we set a resilient threshold u_0 to isolate the background clutter and update the RDM as $\mathcal{R} \leftarrow \text{sgn}\,(u - u_0 J) \circ \mathcal{R}$, where J is a matrix of ones, $\circ$ is Hadamard product [20], and $\text{sgn}\,(\cdot)$ is the sign function which gives 1 when input > 0 and gives 0 for other cases. According to the Eqs. (10) and (11), we can formulate the clutter isolation rate p_c as:

$$p_c = \exp\left[-\exp\frac{\pi}{\sqrt{6}}u_0 - \gamma\right], \tag{13}$$

where γ is the Euler-Mascheroni constant. The above equation indicates that the clutter isolation rate depends on the resilient threshold only, which is a constant false alarm rate (CFAR). Here we set p_c to 10^{-6} and u_0 to 2.5 in the *VocalPrint* accordingly.

5.2 Dynamic Clutter Removal

The moving objects, such as passersby and vehicles, can cause dynamic clutter that could not be removed by applying the resilient threshold on RDM. The reason is that its amplitude of spectrum envelop does not obey the aforementioned Weibull distribution. In this part, we leverage the information across multiple RDMs to remove the dynamic clutter. Specifically, considering that the first FFT performed on each chirp gives the range information (profile) and the second chirp-wise FFT gives the speed information, we utilize the movement of the object to detect and remove the dynamic clutter by making an element-wise comparison among D consecutive RDMs $\mathcal{R}^i (i \in [1, D])$. If $\exists i, j, k : \mathcal{R}^i_{jk} = 0$ is true, the $\mathcal{R}^1_{jk}$ should be regarded as the clutter and we update $\mathcal{R}^1_{jk} \leftarrow 0$. The required number of RDMs to detect the moving objects with velocity v_i is formulated as: $MT_r \frac{1}{D} \sum_i^D v_i \geq \Delta RES$, where M, T_r, and ΔRES are the numbers of range profiles in one RDM, the chirp cycle time, and the range resolution as aforementioned, respectively. Given $D = 16$, the moving objects with average speed $\frac{1}{D} \sum_i^D v_i = 0.11$ m/s that is much slower than the walking speed could be removed.

5.3 Near-body Clutter Mitigation

With the removal of background clutter and dynamic clutter reflected by the moving objects, the mmWave signal mainly includes vocal fold biometric information. However, the mmWave reflected

by the near-body object will still interfere with the phase estimation formulated in Eq. (5) because the reflected clutter is within the same range bin of the vocal vibration.

To mitigate the near-body clutter, we denote the composite amplitude and the initial phase of all the near-body clutter as A_0 and θ_0, respectively. Then, the phasor scatters on the complex phasor diagram can be formulated by the IF signal:

$$H(mT_r + nT_s) = A_0 \exp(j\theta_0) + \sum_{k=1}^{K} A_{m,n}^k \exp(j\theta_{m,n}^k), \tag{14}$$

where T_s is the sampling time interval, $A_{m,n}^k$ and $\theta_{m,n}^k$ represent the amplitude and the initial phase of the k-th tone in the vocal vibration at $mT_r + nT_s$. Considering that the chirp cycle time T_r is far less than the duration of one phoneme, we can rewrite Eq. (14) as:

$$H(mT_r + nT_s) = A_0 \exp(j\theta_0) + \bar{A} \exp(j\theta_m). \tag{15}$$

We estimate the phasor amplitude $\widetilde{A}_m$ by maximizing the likelihood:

$$\widetilde{A}_m = |\frac{1}{N} \sum_{n=0}^{N-1} H(mT_r + nT_s) \exp(-j\omega_H n T_s)|. \tag{16}$$

According to Eq. (15), the phasor scatter set $\mathcal{S} = \{\widetilde{A}_m \angle \psi_m\}, (m = 1, 2, \cdots, M)$ in the phasor diagram satisfies:

$$\|\widetilde{A}_m \angle \psi_m - A_0 \angle \theta_0\|_2 = \bar{A}. \tag{17}$$

Finding the composite amplitude A_0 and initial phase θ_0 of the clutter is equivalent to solving the following optimization problem:

$$\min_{A_0, \angle\psi_0, \bar{A}} \sum_{m=0}^{M} \{\|\widetilde{A}_m \angle \psi_m - A_0 \angle \theta_0\|_2^2 - \bar{A}^2\}. \tag{18}$$

By denoting y as $[2A_0 \cos\theta_0, 2A_0 \sin\theta_0, \bar{A}^2 - oo^\top]$, the closed-form solution to the above minimization problem can be written as $y = (G^\top G)^{-1} G^\top d$, where $G^\top = [g_m]$ is a $3 \times M$ matrix with each column $g_m = [a_m \cos\psi_m, a_m \sin\psi_m, 1]^\top$, and $d = [\|\widetilde{A}_m \angle \psi_m\|_2^2]$, $m \in [1, M]$.

Finally, the near-body clutter mitigation is performed by removing the component $A_0 e^{j\theta_0}$ from Eq. (14) and updating the ψ_m as:

$$\psi_m = \arctan \frac{a_m \sin\psi_m - A_0 \sin\theta_0}{a_m \cos\psi_m - A_0 \cos\theta_0}. \tag{19}$$

Preliminary Results: In our preliminary experiment, we collect RF voice data that contains both static and dynamic clutters and verify our proposed model-centric clutter suppression scheme. Two subjects are asked to sit towards mmWave probe in an uncontrolled outdoor environment with moving backgrounds (*e.g.*, vehicles, and passersby), and pronounce "Ahhh" for around 5 seconds. Other experiment settings are the same as we mentioned in Section 2.2.

- **RDM analysis:** Figure 7 shows the RDMs before and after leveraging clutter suppression scheme, and we observe that the dynamic and background clutters shown in Figure 7(a) are all removed in Figure 7(b). Note that, since near-body clutter is within the same range bin of target vocal vibration signal, the mitigation effect cannot be observed from RDM. The results indicate that our proposed model-centric signal processing scheme is able to mitigate the impact of clutters in RF streaming and obtain precise vocal vibration data.
- **Spectral features analysis:** We further extract spectral centroid and crest from precise vocal vibration signals. The spectral centroid is the indication of the center of gravity of the spectrum, so it can locate large peaks corresponding to approximate formants' positions and pitch frequencies. The spectral crest represents the peakiness of the spectrum that can be used for quantifying the tonality of the signal. They are both typical spectral descriptors that can discriminate between different speakers. As shown in Figure 8, we observe that both spectral centroid and crest possess a great difference between these two subjects in terms of local extremum, mean value, and variation trend. The results indicate that fine-grained spectral proprieties in vocal patterns are well preserved in RF voice data with the help of clutter suppression and can be used for identification.

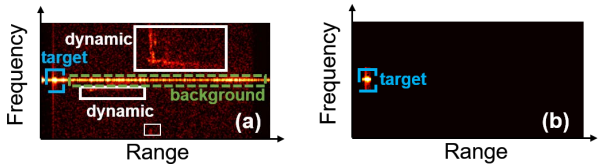

Figure 7: The RDM before (a) and after (b) leveraging clutter suppression scheme.

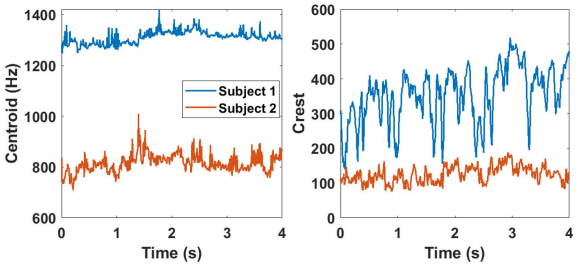

Figure 8: Spectral centroid and crest that are extracted from two subjects' vocal vibration data after clutter suppression.

6 VOCAL AUTHENTICATION

In this section, we explore and identify the optimal biometric features that characterize vocal source and vocal tract information (shown in Figure 9) and input them to an ensemble classifier for robust user authentication.

6.1 Vocal Biometric Features Extraction

Vocal Source Features: The vocal source signal characterizes the muscle structure and tension of the vocal cords, and the related glottal pulse parameters, *e.g.*, closing instants rate, opening duration, and opening degree of the glottis [17]. The vibration pattern of the vocal cords not only provides a voicing source for speech production but also characterizes unique nonlinear flow patterns. The resulting periodic pulse-like epoch shape varies among speakers. Therefore, features derived from the vocal source provide unique physiological information for user identification.

To extract glottal flow cepstrum coefficients (GFCC) that represent the spectral magnitude characteristics of a speaker's glottal excitation pattern, we use the iterative adaptive inverse filtering (IAIF) method to estimate the glottal waveform of speech signal and then perform mel-spaced cepstral analysis [49]. We also derive residual phase Cepstrum coefficients (RPCC) [66] to characterize the phase information of the underlying excitation waveform. Moreover, to

measure the underlying energy required for speech production, we compute the Teager phase cepstrum coefficients (TPCC) [47] that capture phase characteristics of the Teager nonlinear energy model of the speech production [14]. The process for the extraction is two-stepped. First, we apply the Teager-Kaiser energy operator to a band-pass filtered speech signal for calculating excitation energy contour and perform the Hilbert transformation to acquire a fine energy structure. Second, the cepstrum of the fine energy structure is computed and warped to the Mel frequency scale followed by a log and discrete cosine transform (DCT) operation to obtain TPCC.

Vocal Tract Features: Vocal tract system that consists of multiple articulatory organs (*e.g.*, epiglottis, corniculate cartilage, cuneiform cartilage) works as a filter to resonate and reshape vocal source signals. The motion of relevant articulatory organs generates associated articulatory gestures for the flow, but the movement speed and intensity vary from person to person. Therefore, we extract vocal tract features for speaker identification.

We first derive some spectral features, *i.e.*, centroid, band energy, crest, flatness, entropy as the descriptors of the short-term spectral envelope [21], which are the acoustic correlate of timbre. Since coefficients on a linear/nonlinear Mel-scale of frequency can characterize the spectral envelope of a quasi-stationary signal segment, we further extract Mel frequency cepstral coefficients (MFCC) [42] to reflect the resonance properties of the vocal tract system. Specifically, we first convert pre-processed RF vocal biometric signals into a set of mel-frequency spectrums and then employ Triangular band-pass filters to make the signals adhere to the attenuation characteristics of the Mel scale. After the logarithmic compression and DCT, 12-dimensional MFCCs is acquired. To complement MFCCs, linear predictive coefficients (LPC) [53] is selected to characterize formants, *i.e.*, a resonance frequency of the vocal tract. We adopt linear prediction methods to infer the filter coefficients equivalent to the vocal tract by minimizing the mean square error between the input vocal signals and estimated vocal signals. Based on the extracted LPC, we deduce linear predictive cepstral coefficients (LPCC) [31] by performing Cepstral analysis on LPC calculated spectral envelope. We also derive line spectral frequencies (LSF) [40] from LPC, since it can characterize bandwidths and resonance locations and emphasize the spectral peak location.

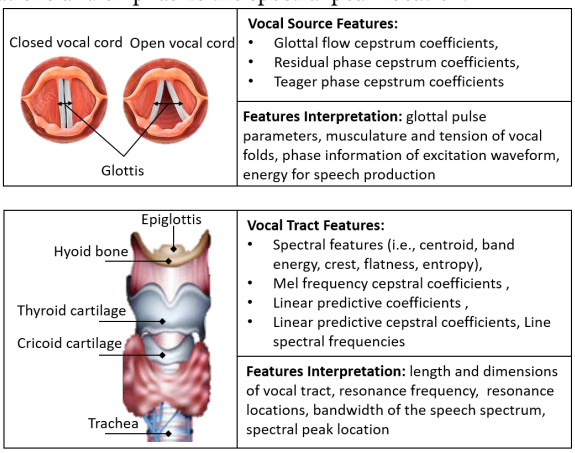

Figure 9: Biometric vocal features and interpretation.

6.2 Fine-tuned Authentication

Biometric feature selection: In practice, not all extracted features are unique enough to distinguish different speakers. Therefore, we use the Fisher Score [56] to select features that are more distinct between classes and consistent within one class. As conventional Fisher Score based selection method processes features individually, and thereby missing high-score feature subsets, we employ a cutting plane algorithm [25] to select a subset of features simultaneously. In each iteration, multivariate ridge regression and projected gradient descent are adopted alternatively to solve a multiple kernel learning problem [52]. After the feature selection, the initial vocal biometric feature vector is reduced to 39 descriptors and then fed to the classification model.

Classifiers Fusion: As the first exploratory study to derive vocal biometric traits from skin-reflected mmWave, we employ the following widely used speaker identification classifiers:

- **Gaussian Mixture Model-Universal Background Model (GMM-UBM):** The use of GMM for modeling speaker identity is because the Gaussian components approximate spectral features and Gaussian mixtures can model arbitrary densities [43]. To guarantee reliable system performance without increasing model complexity, we further introduce UBM to help develop the speaker identification model [39]. The UBM model is trained with expectation-maximization (EM) algorithm on a large amount of data gathered from the background population (*i.e.*, the NIST 2001 one-speaker detection database [38]), then the target speaker model is adapted from the UBM model utilizing training data based on maximum a posteriori (MAP) principle [39]. We calculate the difference of log-likelihood between the target speaker model and the UBM model to determine whether selected features are originated from the genuine speaker or not.
- **Support Vector Machine (SVM):** It is a classification and regression method based on statistical learning theory [13]. We adopt SVM in speaker identification because it can achieve superior generalization performance in classifying unseen data [9]. With the help of kernel functions, the SVM optimizer can find a maximum-margin hyperplane that separates training samples from the genuine speaker and impostor subjects.
- **Hidden Markov Model (HMM)**: It is a statistical tool that describes a Markov process with unobserved states. We select HMM for speaker identification because the states of an HMM characterize the vocal configuration of a speaker and the changes of vocal configuration may duplicate in pronunciation [16]. We use the Baum-Welch algorithm [3] to determine the parameters of an HMM. Then, speaker identification is performed by a Viterbi algorithm to compute likelihood scores for each signal [15].

Finally, we combine the output scores of these three classifiers by weighted sum and optimize the fusion weights based on logistic regression. The BOSARIS Toolkit [6] is employed for implementing fusion and determining the genuine speaker.

7 SYSTEM IMPLEMENTATION AND EVALUATION SETUP

7.1 System Implementation

The selection criteria for mmWave probe hardware depends upon the desired waveform characteristics that take two major factors

into consideration. First, the chirps and frames generated by the probe should be able to capture the high-resolution vocal vibration from the target user. Second, the mmWave probe should guarantee the minimum signal-to-noise ratio (SNR) so that the proposed data processing techniques can distinguish among vocal vibrations and the clutter. Therefore, we carefully design the mmWave waveform configuration as shown in Table 1. This configuration enables the range resolution of 3.75 cm, displacement resolution around 1 mm [57], which satisfies the requirements for sensing the vocal vibrations. Our design of the mmWave waveform can be generated effortlessly by any off-the-shelf mmWave probe [44, 45] or customized hardware [19], thereby facilitating affordability (less than $70), portability (100g) and energy-efficiency (135mW) in real-world setups. In our work, we leverage a Texas Instruments AWR1642 mmWave probe (TX Power=12.5 dBm, RX Gain=30 dB) [62] to emit the signal and capture the data. The range profiles are generated on-board and then transferred to the laptop for further processing.

Table 1: mmWave waveform design.

Frequency Slope	71.5 MHz/μs	Bandwidth	4 GHz
ADC Samples/Second	5000K	Idle Time	10 μs
Chirp Cycle Time	65.8 μs	Chirps/Frame	128
Frame Periodicity	9 ms	Samples/Chirp	256

7.2 Evaluation setup

Experiment preparation. We conduct extensive experiments to confirm the capability of *VocalPrint* for user authentication. Figure 10 shows the experimental setup. A subject is asked to sit in a chair. We align the customized mmWave probe in the direction of the subject's throat, *i.e.* the subject orientation is 0° with respect to the probe. The distance between the subject and the probe is 20cm. All subjects are required to sit in the same position unless specified in the evaluation. The mmWave probe connects to a 5.0 V power supply, and the working current is 2A. We deploy two laptops with the Windows 10 operating system. One is used for collecting the signals from the receiving terminal of a mmWave probe, using the network interface card. The other is employed to display the reading materials. The training processes are done by a workstation equipped with an Intel Xeon E5-1620 v4 @ 3.50GHz.

Data collection. Our biometric study is approved by IRB. 41 subjects (21 males and 20 females) are asked to read *The North wind and the sun passage* (113 words) and the first two sentences of *The Grandfather Passage* (37 words) following a prompter to guarantee the same reading time. On average, each subject takes around 51 seconds for *The North wind and the sun* and 14.6 seconds for the first two sentences of *The Grandfather Passage*. The collected data are anonymous and stored locally to protect the subject's privacy.

Partition. To evaluate the performance with text-independent features, we use the received signals corresponding to *The north wind and the sun* for training and *The Grandfather Passage* for testing. The received signals are segmented evenly with a 50% overlapping rate and then filtered by an efficient speech detection mechanism based on the Zero Cross Rate and Root Mean Square in time domain [35] to isolate non-speech segments. The segment lengths are varied as 5ms, 10ms, 15ms, 20ms, 25ms, and 30 ms, respectively, for performance analysis. Based on segment length, we finally collect 111720–672000 samples and use 71400–428400 samples for training and the rest for testing. Among the overall 41 subjects, each acts as a genuine user once while the remaining 40 subjects act as imposters to access the system. Therefore, the genuine subjects and imposters ratio is 1:40 for every authentication trial.

Evaluation metrics. We introduce F-score, balanced accuracy (BAC), receiver operating characteristics (ROC) curve, equal error rate (EER) as metrics in our evaluation since these are non-sensitive to class distribution for evaluating authentication systems [34, 80].

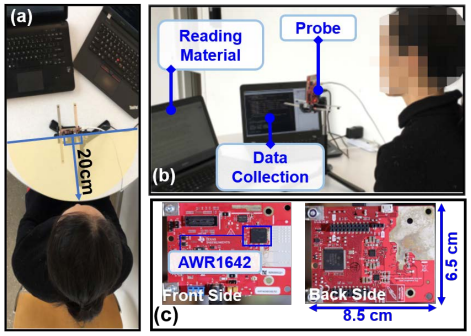

Figure 10: The evaluation setup: (a) subject's sitting position; (b) in a lab environment; (c) mmWave probe architecture.

8 PERFORMANCE EVALUATION

In this section, we evaluate the performance and robustness of *VocalPrint* for authentication. All the results are obtained after the body motion compensation except the one specified as "before body motion compensation" in the evaluation of subjects in motion.

8.1 Overall System Performance

To maximize the applicability of *VocalPrint* in real-world scenarios, it is important that the system can not only differentiate between the legitimate users and imposters but also perform the authentication in a timely fashion. The authentication time is defined as the total time elapsed to make a final prediction and is dependent on the segment length needed to authenticate users. To determine an optimal segment length of mmWave signal for precise authentication, we evaluate the system performance with segment lengths as 5ms, 10ms, 15ms, 20ms, 25ms, 30ms, respectively.

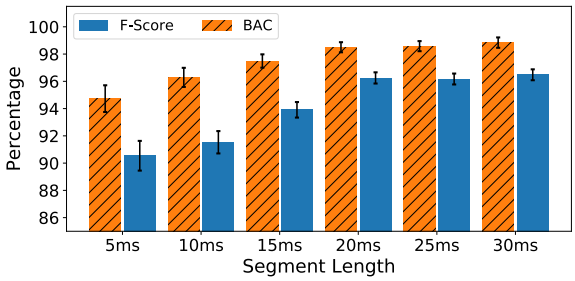

Figure 11: The overall performance of *VocalPrint* with different segment lengths.

Figure 11 illustrates the F-score and BAC measure for 41 subjects with different segment lengths. We observe that when the segment length is less than 15ms, it does not contain sufficient information for accurate authentication, indicated by the low BAC and F-score, and high standard deviation (STD). The performance is improved when the length of the segment is increased, however, the improvement in F-score and BAC is not significant after the segment length

is increased from 20ms to 30ms. Specifically, BAC achieves 98.52%, 98.58%, and 98.85% with the STD of 0.37%, 0.37% and 0.38% for 20ms, 25ms and 30ms, respectively. F-score reaches 96.27%, 96.18%, and 96.46% with the STD of 0.41%, 0.4% and 0.4% for 20ms, 25ms and 30ms.

For a more concrete analysis, we also plot the ROC curves and calculate the corresponding area-under-curve (AUC) with different segment lengths, as shown in Figure 12. Although the 30ms segment achieves the best performance, the performance is not improved significantly compared with the 20ms segment. The corresponding EER are given as 9.91%, 10.18%, 9.08%, 4.97%, 4.99%, and 4.92%, respectively. These results are consistent with BAC and F-score.

Based on the above observations, we conclude that the segment length of 20ms is most appropriate for training and testing. With such a segment length, the total time needed to verify a user is 340ms. The results also demonstrate the effectiveness of *VocalPrint* for reliable user authentication. For the remainder of this paper, we use the segment length of 20ms during the performance analysis.

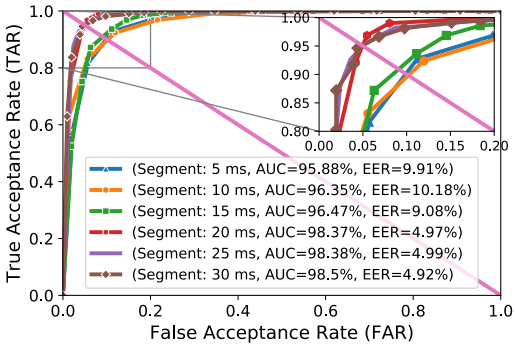

Figure 12: The ROC and EER with different segment lengths.

8.2 Robustness Analysis

Impact of variant distances and orientations. To maximize the user experience in real practice, *VocalPrint* should be tolerant to variations in the sensing position. As a result, we study whether the performance of *VocalPrint* will be affected by changeable distances and orientations. In the experiment, the subjects are asked to read the first two sentences of *The Grandfather Passage* while the orientation and sensing distance between the subject's throat and mmWave probe is varied from 0° to 60° and 0.2m to 2m, respectively. The results are shown in Figure 13. We observe that the BAC reaches up to 96% when the sensing distance is less than 150 cm and human orientation is within 45°. Within 0.5m, the *VocalPrint* can still achieve above 96% as human orientation expands to 60°.

Due to mmWave attenuation, the *VocalPrint* performance drops as the sensing distance increases. Therefore, we want to further explore *VocalPrint* can work in what kind of application scenarios and at which level of authentication accuracy, when extending the distance. Specifically, we evaluate *VocalPrint* performance in subdivided daily-life scenarios: 1) body field (0-0.5m): communication with smartphone and wearable device; 2) social distancing field (0.5m-2m): interaction with car and desktop device; 3) local field (2m-5m): interaction with the smart home appliance. Figure 14 shows that *VocalPrint* can achieve over 98% BAC in the body field

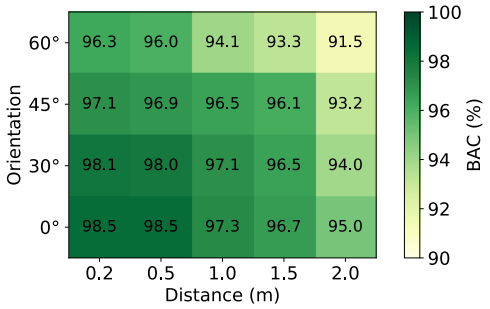

Figure 13: The performance of *VocalPrint* under different sensing distances and human orientations.

and over 95% BAC in the social distancing field. As the distance increases to 460 cm, the authentication accuracy is around 91.7%.

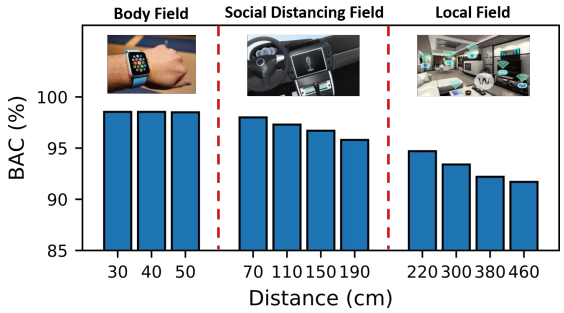

Figure 14: The performance of *VocalPrint* in the body field, social distancing field, and local field.

Impact of body posture and motion. To enhance usability, *VocalPrint* should facilitate accurate user authentication at all times, without requiring users to stop their ongoing activities (*e.g.,* driving) or put down any object held in their other hand. Therefore, we investigate the performance of *VocalPrint* under the effect of body posture and motion. Specifically, we study (1) posture and motion that may shelter the near-throat skin surface from mmWave sensing; (2) periodic body posture and motions. In our experiment, while reading the first two sentences of *The Grandfather Passage*, each subject is asked to continuously perform four common daily-life activities, including rhythmic movements during listening to music, combing hair, mimic driving, and writing, thereby exhibiting minute to large-scale body motion.

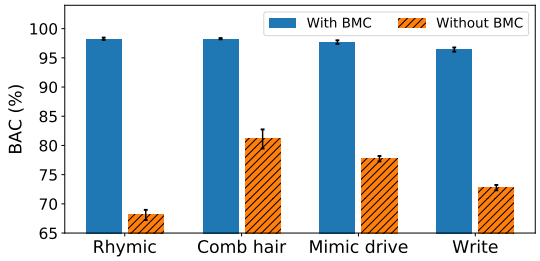

Figure 15: The BAC comparison with and without body motion compensation.

With the current experimental setting, the BAC results before and after body motion compensation are shown in Figure 15. With the body motion compensation model, we observe that the BAC corresponding to the subject performing a rhythmical movement

and combing hair during authentication reaches above 98%, while the BAC for sheltering motions (*i.e.,* writing and mimic driving) varies between 96% and 98%. The results demonstrate that body motion resulting in sheltering of the near-throat skin surface from mmWave sensing can affect system performance to some extent. This is due to the limited penetration capability of 77GHz mmWave that is leveraged in this work [74]. Meanwhile, without the body motion compensation method proposed in our work, the BAC gets reduced to an average performance of 73%. Regardless of body motion, *VocalPrint* shows a reliable performance in user authentication.

Impact of wearable accessories. In our daily life, it is common for users to wear accessories around the throat. Therefore, we are motivated to examine whether the authentication performance will be affected by the wearable accessories which are made of different materials (e.g., metal, cotton, wool, plastic) and pose partial or full occlusion to the throat. Specifically, the subjects are asked to wear necklaces, shirt collar, scarf, and earbuds, respectively while reading the first two sentences of *The Grandfather Passage*. Figure 16 shows that *VocalPrint* achieves more than 98% BAC with necklace, shirt collar, and wool scarf around neck, and 97.7% BAC with earbuds. Therefore, *VocalPrint* is robust to wearable accessories.

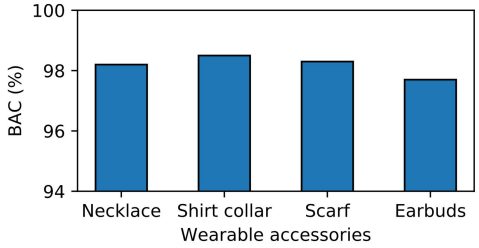

Figure 16: The performance of *VocalPrint* with different wearable accessories.

Impact of speaking content and speed. We also conduct experiments to test the impact of speaking content and speed on the *VocalPrint* performance. To set different speaking speed, we ask 10 subjects to read the same sentences (37 words) for the test in a slow (completion time is around 20s), normal (completion time is around 15s), fast speed (completion time is around 10s), respectively. Figure 17 (a) shows that the average BAC values are between 98.1% and 98.6% when the speaking speed varies. To evaluate the impact of speech content, the subjects are asked to read three different sentences from"The Rainbow Passage", "Comma Gets a Cure", and "Arthur the Rat" at a normal speed for the test. As illustrated in Figure 17 (b), *vocalPrint* can achieve around 98.5% BAC with different reading materials. The results indicate that *VocalPrint* is robust to speaking speed and content because we extract intrinsic vocal source and tract information for the training model.

Longitudinal Study. For any biometric method, permanence is a critical factor. We examine the permanence of vocal vibrations to show the potential of *VocalPrint* as an enhancement to voice authentication. 20 subjects (10 males and 10 females) participate in the long-term study lasting 30 days. In every period of three days, each subject reads the first two sentences of *The Grandfather Passage*, and mmWave signals are obtained. The training set is generated based on the collected mmWave signals on the first day of enrollment. As Figure 18 shown, the average values of BAC are

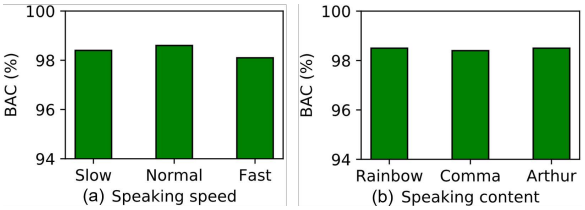

Figure 17: The performance of *VocalPrint* with different speaking speed and content.

between 98% and 99%, and the STDs are between 0.37 and 0.39 in the 30-day duration. We can conclude that there is no notable decreasing and ascending tendency on average BAC results, which indicates that *VocalPrint* is robust to the time change.

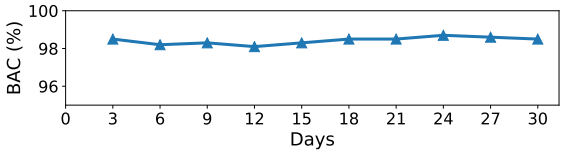

Figure 18: A 30-day longitudinal study.

9 AMBIENT RESILIENCE STUDY

9.1 Acoustic Noise

In practice, there are two primary types of acoustic noise with different spectral characteristics: pop music and presentation. To evaluate the authentication performance in presence of acoustic noise, a loudspeaker is placed next to the user and plays the recorded music and presentation sound at different decibel levels (i.e., volume varies as 0, 25%, 50% and, 75%). At the same time, each subject reads the first two sentences of *The Grandfather Passage*. With the current experiment setting, we obtain the BAC and F-score under different volumes of music and presentation sound, as shown in Figure 19. We observe that the authentication accuracy does not exhibit much difference under music and presentation sound. Even when the loudspeaker volume increases from 0 to 75%, the values of BAC and F-score remain stable. To be specific, the BAC values are between 98% and 99%, and the F-score values are between 96% and 97%. These results validate that *VocalPrint* is immune from different types and volumes of acoustic noises. This is consistent with the fact that *VocalPrint* employs an electromagnetic channel that is not affected by acoustic noise in the ambient environment.

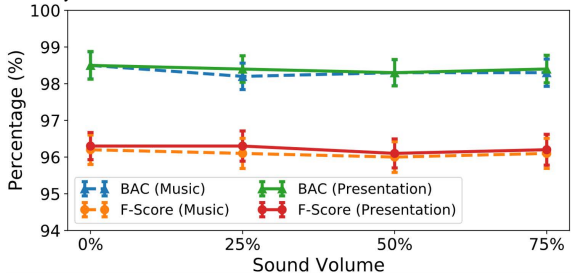

Figure 19: Performance under acoustic noise.

9.2 Environmental Dynamics

It is a known fact that variations in the sensing environment can significantly affect the quality of the received signal, increasing the false acceptance rate of the authentication system. Specifically,

mmWave signals may be affected by the stationary and moving objects in the environment. To evaluate the capability of our proposed resilience-aware suppression model, we select three ambient conditions: (1) snowy outdoor with environmental temperature as −5°C (23°F) and no human obstruction; (2) student lounge with environmental temperature as 20°C (68°F) and periodic human obstruction; (3) three people continuously walking around the mmWave probe within 2m distance, depicting constant human obstruction. The subjects are asked to read the first two sentences of *The Grandfather Passage* in different ambient conditions as mentioned above. Figure 20 shows the authentication results. In each ambient condition, the BAC reaches over 98% with around 0.4% STD, and the F-score values are more than 96% with approximately 0.4% STD. These results indicate that the resilience-aware clutter suppression approach can effectively remove the clutters in the usual authentication scenarios. Therefore, *VocalPrint* is resilient against environmental dynamics and can be applied in real-world scenarios.

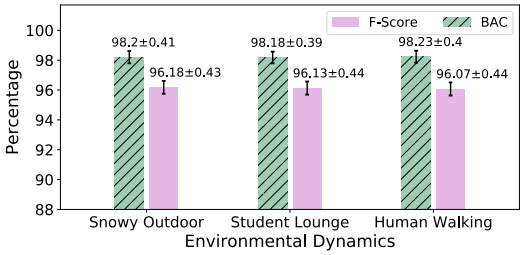

Figure 20: Performance under environmental dynamics.

10 SPOOFING RESILIENCE STUDY

10.1 Counterfeit Attack

We assume that the attacker (1) knows that the uniqueness of human voice is intrinsically sourced in vocal vibration; (2) observes that when a person speaks, vocal cord vibrations caused by air pressure are propagated through the vocal tract and can be measured on the skin surface. Based on this knowledge, the attacker may forge the target's vocal vibration to spoof *VocalPrint*. To verify if a human's vocal vibration can be simulated, we construct a counterfeit attack model, as shown in Figure 21(a). We place an audio transducer inside a throat model to replay a pre-recorded passphrase of the target user. As illustrated in Figure 21(b), the transducer is used to simulate the vocal source excitation signal (*i.e.,* vocal cord vibrations caused by air pressure). When the audio signal passes the coil of the transducer, a dynamic electro-magnetic field is generated, which makes the actuator vibrate the throat model. The supralaryngeal vocal tract in our throat model acts as reshaping the source signal, as shown in Figure 21(d). Finally, the forged vocal vibration is reflected by a readily available bionic skin material (*i.e.,* Silicone [12]) covering the throat model (see Figure 21(c)).

To overcome this counterfeit attack, we implement a body motion detector in the random motion compensation module (see Section 4.3) to judge whether the reflected vocal vibration is originated from a live user or a model. Specifically, the detector examines the value of range shift χ_m in Eq. (8) when the envelop correlation function reaches maximum. If $\chi_m = 0$, it implies that the range bins misalignment issue does not exist, *i.e.*, there is no random body motion. To evaluate the effectiveness, we place the forged throat model

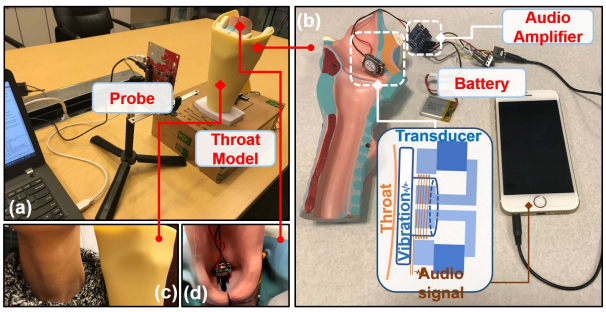

Figure 21: Counterfeit attack experiment setup. The adversary counterfeits vocal vibrations using an audio transducer and 1:1 throat model.

at a distance of 20cm to *VocalPrint*, and try 200 trials. Compared with *CaField* [71] that shows 0.82% (under loudspeaker-based imposters) and 1.87% (under human imposters) false acceptance rate (FAR), *VocalPrint* achieves 1.5% FAR under forged throat attacks.

We also consider a challenging scenario where the audio transducer (*i.e.*, vibration part) is stuck to the actual human throat. We launch 200 attacks, only 9 (4.5%) forged models are misclassified as legitimate users. To summarize, experiment results indicate that *VocalPrint* can combat this counterfeit attack.

10.2 Mimicry Attack

Some attackers may intend to compromise the *VocalPrint* by mimicking the speaker. To verify whether *VocalPrint* can defend against mimicry attack, 10 volunteers are invited to mimic the target speaker. These volunteers are face-to-face with the target speaker and observe how they pronounce speech. After the volunteers repeatedly practice pronunciation by mimicking 1) the articulatory movement of upper and lower lips, tongue and jaw; 2) speaking speed, intonation, rhythm, conversation-level characteristics (*e.g.*, "uh-huh", "oh yeah", etc.) of the speaker, they initiate the mimicry attacks in front of the *VocalPrint*. Each volunteer mimics 5 target speakers for 10 trials. In all, 500 mimicry attacks are launched to *VocalPrint* but every attack fails. The results are as expected because the human voice is individually unique and cannot be entirely mimicked.

10.3 Signal Replay Attack

We assume that the attacker knows the details of the mmWave probe used to sense the vocal vibration. Understanding this fact, the attacker can first eavesdrop on the communication between the speaker and *VocalPrint* to record the skin-reflected signals. After the eavesdropping process, the attacker can deploy devices to absorb emitted mmWave signals, and then replay the pre-recorded skin-reflected signals to spoof *VocalPrint*. To defend this signal replay attack, every time to emit mmWave signals, *VocalPrint* randomly selects three chirps and alters chirp rates, thereby the frequency shift value of the range profile will change after performing FFT on IF signals. In this case, when an attacker replays pre-recorded signals in a new time, such signals can be easily refused by comparing the frequency shift value. To examine the effectiveness, we record the reflected signals from different speakers and use a mmWave signal generator to send the imitation signals to *VocalPrint*. The results show that *VocalPrint* can recognize all the imitation signals.

Table 2: A comparison of voice-based authentication methods.

System	Liveness Detection Principle	Sensing Mechanism	Sensing Orientation	Acoustic Noise Sensitivity	User Cooperation	Test Distance
VAuth [18]	Body vibration	Bone conduction	N/A	Resistive	Yes	Contact
Chen et al. [10]	Magnetic field	Magnetic	Directional	Resistive	Yes	<10 cm
CaField [71]	Sound field	Acoustic	Directional	Sensitive	Yes	<50 cm
VoicePop [67]	Pop noise	Acoustic	Non-directional	Sensitive	Yes	<10 cm
VoiceLive [77]	TDoA of phenome sounds	Acoustic	Non-directional	Sensitive	Yes	<50 cm
VoiceGesture [76]	Mouth motion	Acoustic (ultrasonic)	Non-directional	Sensitive	Yes	<50 cm
WiVo [41]	Mouth motion	Radio frequency	Directional	Resistive	No	<50 cm
VocalPrint	**Vocal vibration**	**Radio frequency**	**Directional**	**Resistive**	**No**	**0 - 200 cm**

11 RELATED WORK

Voice authentication. Voice authentication is a historical topic in biometrics and has been studied well [4, 41]. Existing studies show that most voice authentication solutions are vulnerable against spoofing attacks [22, 51]. To defense attacks, many liveness detection approaches have been proposed based on the distinction between human and loudspeaker [58, 59]. For example, mouth motion in speaking is distinct from the manner that the loudspeaker vibrates the diaphragm. Based on it, some studies leveraged RF reflections [41] and ultrasonic reflections [76] of the mouth motion for liveness detection, but mouth motion is observable and can be mimicked potentially. Other studies exploited characteristics exclusive to human speakers or loudspeakers, such as magnetic field emitted from the loudspeaker [10], pop music in human utterances [67], time-difference-of-arrival from two microphones [77], and sound field [71]. However, these microphone-based solutions are intrinsically sensitive to ambient noise and also assume that replaying cannot generate identical sound waves. By introducing a non-sound based sensing modality, *VocalPrint* can protect the system even if attackers generate identical sound waves. Other solutions leveraged contact-based sensors for voice authentication, such as VAuth [18], Vocal Resonance [35] which can immune ambient noise. However, these bone-conduction solutions require skin-contact and sacrifice usability. According to the literature (see Table 2), no solution exists in addressing all these issues in voice authentication.

mmWave-based human sensing. Recent advances have demonstrated that mmWave accurately detects minute variations caused by a human without body contact [1, 24]. Some works leverage mmWave into activity recognition [23, 32, 79] and emotion recognition [78]. Some other works focus on the detection of biometrics [72, 78]. For example, Petkie et al. [48] employed a 228 GHz heterodyne radar to measure the respiration and heart rates at a distance of 10 meters. Lin and Song *et al.* [34] implemented Cardiac Scan, a non-contact and continuous sensing system for user authentication. Recent workS [33, 69] leverage mmWave to sense voice-related information to facilitate voice-user interface. Compared with these works, our work explores vocal vibration as a continuous and non-contact biometric identification and captures anti-spoofing features (*i.e.*, throat physiological intrinsic) to defend against malicious attacks.

12 LIMITATION

Long sensing distance. As the distance increases, velocity resolution of the range profile will correspondingly degrade, thereby the performance of background clutter isolation will be affected and finally lead to decreased authentication accuracy. To extend effective sensing distance for remote voice biometric-based application, we can increase the bandwidth of IF signals in mmWave waveform design [11].

Sensing orientation. *VocalPrint*'s performance is affected by the user orientation especially in long-distance scenarios. To enhance the robustness of *VocalPrint*, one possible solution is to collect user's vocal vibration from different orientations with respect to the probe in the enrollment phase, because vocal sounds travel through the bone and the resulting vibration on the neck surface (besides the throat region) is also valuable for authentication [35]. This is similar to the enrollment phase of FaceID which requires the user to move his/her head slowly to complete a circle [7].

Compatibility with IoT Devices. *VocalPrint* employs a high sampling rate to capture fine-grained vocal vibration, but such a high sampling rate will produce massive data samples. The future mmWave-enabled smart device is possible to implement signal processing based resilience-aware clutter suppression and vocal authentication in a real-time fashion with the help of the high-speed DSP [64].

13 CONCLUSION

Existing voice authentication systems are vulnerable to noise interference and spoof attacks. In this paper, we introduce a novel biometric system, *VocalPrint*, for resilient security of voice authentication. Specifically, *VocalPrint* is on the basis of a 77GHz FMCW probe to sense the minute vocal vibrations in near-throat region of users and leverage the skin-reflect mmWave signals. A novel resilience-aware clutter suppression approach is proposed to isolate the complex ambient noise and body motion from the mmWave signals and allow further extraction of unique vocal tract and vocal source features. Extensive experiments indicate that the authentication accuracy of *VocalPrint* exceeds 96% even under unfavorable conditions. We also show the ambient resilience and spoof resilience of *VocalPrint* to show its practicality in real-world setups. In future work, we plan to evaluate *VocalPrint* with more people suffering from speech disorders and improve system accuracy.

ACKNOWLEDGMENTS

We thank all anonymous reviewers for their insightful comments on this paper. This work was supported by the National Science Foundation under grant No. ECCS-2028872, CNS-1718375.

REFERENCES

[1] Fadel Adib, Hongzi Mao, Zachary Kabelac, Dina Katabi, and Robert C Miller. 2015. Smart homes that monitor breathing and heart rate. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. ACM, 837–846.

[2] M. L. Attiah, M. Ismail, R. Nordin, and N. F. Abdullah. 2015. Dynamic multi-state ultra-wideband mm-wave frequency selection for 5G communication. In *2015 IEEE 12th Malaysia International Conference on Communications (MICC)*. 219–224. https://doi.org/10.1109/MICC.2015.7725437

[3] Leonard E Baum and John Alonzo Eagon. 1967. An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. *Bull. Amer. Math. Soc.* 73, 3 (1967), 360–363.

[4] Logan Blue, Hadi Abdullah, Luis Vargas, and Patrick Traynor. 2018. 2ma: Verifying voice commands via two microphone authentication. In *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*. ACM, 89–100.

[5] Rudolf Maarten Bolle, Jonathan Hudson Connell, and Nalini K Ratha. 2005. System and method for liveness authentication using an augmented challenge/response scheme. US Patent 6,851,051.

[6] Niko Brümmer and Edward De Villiers. 2013. The bosaris toolkit: Theory, algorithms and code for surviving the new dcf. *arXiv preprint arXiv:1304.2865* (2013).

[7] Andrew Bud. 2018. Facing the future: The impact of Apple FaceID. *Biometric Technology Today* 2018, 1 (2018), 5–7.

[8] Joseph P Campbell. 1997. Speaker recognition: A tutorial. *Proc. IEEE* 85, 9 (1997), 1437–1462.

[9] William M Campbell, Joseph P Campbell, Douglas A Reynolds, Elliot Singer, and Pedro A Torres-Carrasquillo. 2006. Support vector machines for speaker and language recognition. *Computer Speech & Language* 20, 2-3 (2006), 210–229.

[10] Si Chen, Kui Ren, Sixu Piao, Cong Wang, Qian Wang, Jian Weng, Lu Su, and Aziz Mohaisen. 2017. You can hear but you cannot steal: Defending against voice impersonation attacks on smartphones. In *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 183–195.

[11] Jae-Hyun Choi, Jong-Hun Jang, and Jin-Eep Roh. 2015. Design of an FMCW radar altimeter for wide-range and low measurement error. *IEEE Transactions on Instrumentation and Measurement* 64, 12 (2015), 3517–3525.

[12] Tarang Chugh, Kai Cao, and Anil K Jain. 2018. Fingerprint spoof buster: Use of minutiae-centered patches. *IEEE Transactions on Information Forensics and Security* 13, 9 (2018), 2190–2202.

[13] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20, 3 (1995), 273–297.

[14] Sharmistha Das and John HL Hansen. 2004. Detection of voice onset time (VOT) for unvoived stops (/p/,/t/,/k/) using the Teager energy operator (TEO) for automatic detection of accented English. In *Proceedings of the 6th Nordic Signal Processing Symposium, 2004. NORSIG 2004.* Citeseer, 344–347.

[15] TK Das and KM Nahar. 2016. A voice identification system using hidden markov model. *Indian Journal of Science and Technology* 9, 4 (2016).

[16] Mangesh S Deshpande and Raghunath S Holambe. 2008. Text-independent speaker identification using hidden Markov models. In *2008 First International Conference on Emerging Trends in Engineering and Technology*. IEEE, 641–644.

[17] Gunnar Fant. 1970. *Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations.* Number 2. Walter de Gruyter.

[18] Huan Feng, Kassem Fawaz, and Kang G Shin. 2017. Continuous authentication for voice assistants. In *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*. 343–355.

[19] J. Hasch, E. Topak, R. Schnabel, T. Zwick, R. Weigel, and C. Waldschmidt. 2012. Millimeter-Wave Technology for Automotive Radar Sensors in the 77 GHz Frequency Band. *IEEE Transactions on Microwave Theory and Techniques* 60, 3 (March 2012), 845–860. https://doi.org/10.1109/TMTT.2011.2178427

[20] Roger A Horn. 1990. The hadamard product. In *Proc. Symp. Appl. Math*, Vol. 40. 87–169.

[21] Danoush Hosseinzadeh and Sridhar Krishnan. 2007. Combining vocal source and MFCC features for enhanced speaker recognition performance using GMMs. In *2007 IEEE 9th Workshop on Multimedia Signal Processing*. IEEE, 365–368.

[22] Artur Janicki, Federico Alegre, and Nicholas Evans. 2016. An assessment of automatic speaker verification vulnerabilities to replay spoofing attacks. *Security and Communication Networks* 9, 15 (2016), 3030–3044.

[23] Wenjun Jiang, Chenglin Miao, Fenglong Ma, Shuochao Yao, Yaqing Wang, Ye Yuan, Hongfei Xue, Chen Song, Xin Ma, Dimitrios Koutsonikolas, et al. 2018. Towards Environment Independent Device Free Human Activity Recognition. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*. ACM, 289–304.

[24] Ossi Johannes Kaltiokallio, Hüseyin Yigitler, Riku Jäntti, and Neal Patwari. 2014. Non-invasive respiration rate monitoring using a single COTS TX-RX pair. In *Proceedings of the 13th international symposium on Information processing in sensor networks*. IEEE Press, 59–70.

[25] James E Kelley, Jr. 1960. The cutting-plane method for solving convex programs. *Journal of the society for Industrial and Applied Mathematics* 8, 4 (1960), 703–712.

[26] Lawrence George Kersta. 1962. Voiceprint identification. *Nature* 196, 4861 (1962), 1253–1257.

[27] Bernd J Kröger, Georg Schröder, and Claudia Opgen-Rhein. 1995. A gesture-based dynamic model describing articulatory movement data. *The Journal of the Acoustical Society of America* 98, 4 (1995), 1878–1889.

[28] Jeffrey C Lagarias, James A Reeds, Margaret H Wright, and Paul E Wright. 1998. Convergence properties of the Nelder–Mead simplex method in low dimensions. *SIAM Journal on optimization* 9, 1 (1998), 112–147.

[29] Selena Larson. 2017. Google Home now recognizes your individual voice. *CNN Money, San Francisco, California* 3 (2017).

[30] Changzhi Li, Victor M Lubecke, Olga Boric-Lubecke, and Jenshan Lin. 2013. A review on recent advances in Doppler radar sensors for noncontact healthcare monitoring. *IEEE Transactions on microwave theory and techniques* 61, 5 (2013), 2046–2060.

[31] Penghua Li, Fangchao Hu, Yinguo Li, and Yang Xu. 2014. Speaker identification using linear predictive cepstral coefficients and general regression neural network. In *Proceedings of the 33rd Chinese Control Conference*. IEEE, 4952–4956.

[32] Jaime Lien, Nicholas Gillian, M Emre Karagozler, Patrick Amihood, Carsten Schwesig, Erik Olson, Hakim Raja, and Ivan Poupyrev. 2016. Soli: Ubiquitous gesture sensing with millimeter wave radar. *ACM Transactions on Graphics (TOG)* 35, 4 (2016), 142.

[33] C. Lin, S. Chang, C. Chang, and C. Lin. 2010. Microwave Human Vocal Vibration Signal Detection Based on Doppler Radar Technology. *IEEE Transactions on Microwave Theory and Techniques* 58, 8 (Aug 2010), 2299–2306. https://doi.org/10.1109/TMTT.2010.2052968

[34] Feng Lin, Chen Song, Yan Zhuang, Wenyao Xu, Changzhi Li, and Kui Ren. 2017. Cardiac scan: A non-contact and continuous heart-based user authentication system. In *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*. ACM, 315–328.

[35] Rui Liu, Cory Cornelius, Reza Rawassizadeh, Ronald Peterson, and David Kotz. 2018. Vocal resonance: Using internal body voice for wearable authentication. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 19.

[36] Bram Lohman, Olga Boric-Lubecke, VM Lubecke, PW Ong, and MM Sondhi. 2002. A digital signal processor for Doppler radar sensing of vital signs. *IEEE Engineering in Medicine and Biology Magazine* 21, 5 (2002), 161–164.

[37] Judith A Markowitz. 2000. Voice biometrics. *Commun. ACM* 43, 9 (2000), 66–73.

[38] Alvin F Martin and Mark A Przybocki. 2001. The NIST speaker recognition evaluations: 1996-2001. In *2001: A Speaker Odyssey-The Speaker Recognition Workshop*.

[39] Jack McLaughlin, Douglas A Reynolds, and Terry Gleason. 1999. A study of computation speed-ups of the GMM-UBM speaker recognition system. In *Sixth European Conference on Speech Communication and Technology*.

[40] Ian Vince McLoughlin. 2008. Line spectral pairs. *Signal processing* 88, 3 (2008), 448–467.

[41] Yan Meng, Zichang Wang, Wei Zhang, Peilin Wu, Haojin Zhu, Xiaohui Liang, and Yao Liu. 2018. WiVo: Enhancing the Security of Voice Control System via Wireless Signal in IoT Environment. In *Proceedings of the Eighteenth ACM International Symposium on Mobile Ad Hoc Networking and Computing*. ACM, 81–90.

[42] K Sri Rama Murty and Bayya Yegnanarayana. 2005. Combining evidence from residual phase and MFCC features for speaker recognition. *IEEE signal processing letters* 13, 1 (2005), 52–55.

[43] Seiichi Nakagawa, Kouhei Asakawa, and Longbiao Wang. 2007. Speaker recognition by combining MFCC and phase information. In *Eighth annual conference of the international speech communication association*.

[44] National Instruments [n.d.]. mmWave Transceiver System. http://www.ni.com/sdr/mmwave/

[45] NXP [n.d.]. S32R27 Reference Design Kit for high-performance Automotive Radar. https://www.nxp.com/products/power-management/system-basis-chips/functional-safety-sbcs/s32r27-reference-design-kit-for-high-performance-automotive-radar:RDK-S32R274

[46] J. D. Park and W. J. Kim. 2006. An Efficient Method of Eliminating the Range Ambiguity for a Low-Cost FMCW Radar Using VCO Tuning Characteristics. *IEEE Transactions on Microwave Theory and Techniques* 54, 10 (Oct 2006), 3623–3629. https://doi.org/10.1109/TMTT.2006.882869

[47] Hemant A Patil and Pallavi N Baljekar. 2012. Classification of normal and pathological voices using TEO phase and Mel cepstral features. In *2012 International Conference on Signal Processing and Communications (SPCOM)*. IEEE, 1–5.

[48] Douglas T Petkie, Erik Bryan, Carla Benton, and Brian D Rigling. 2009. Millimeter-wave radar systems for biometric applications. In *Millimetre Wave and Terahertz Sensors and Technology II*, Vol. 7485. International Society for Optics and Photonics, 748502.

[49] Michael David Plumpe, Thomas F Quatieri, and Douglas A Reynolds. 1999. Modeling of the glottal flow derivative waveform with application to speaker identification. *IEEE Transactions on Speech and Audio Processing* 7, 5 (1999), 569–586.

[50] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society.

[51] Jianwei Qian, Haohua Du, Jiahui Hou, Linlin Chen, Taeho Jung, and Xiang-Yang Li. 2018. Hidebehind: Enjoy Voice Input with Voiceprint Unclonability and Anonymity. In *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*. ACM, 82–94.

[52] Alain Rakotomamonjy, Francis Bach, Stéphane Canu, and Yves Grandvalet. 2007. More efficiency in multiple kernel learning. In *Proceedings of the 24th international conference on Machine learning*. 775–782.

[53] Ravi P Ramachandran, Mihailo S Zilovic, and Richard J Mammone. 1995. A comparative study of robust linear predictive analysis methods with applications to speaker identification. *IEEE transactions on speech and audio processing* 3, 2 (1995), 117–125.

[54] Douglas A Reynolds and Richard C Rose. 1995. Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE transactions on speech and audio processing* 3, 1 (1995), 72–83.

[55] Nirupam Roy and Romit Roy Choudhury. 2016. Listening through a vibration motor. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 57–69.

[56] Syed Muhammad Saqlain, Muhammad Sher, Faiz Ali Shah, Imran Khan, Muhammad Usman Ashraf, Muhammad Awais, and Anwar Ghani. 2019. Fisher score and Matthews correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines. *Knowledge and Information Systems* 58, 1 (2019), 139–167.

[57] S. Scherr, S. Ayhan, B. Fischbach, A. Bhutani, M. Pauli, and T. Zwick. 2015. An Efficient Frequency and Phase Estimation Algorithm With CRB Performance for FMCW Radar Applications. *IEEE Transactions on Instrumentation and Measurement* 64, 7 (July 2015), 1868–1875. https://doi.org/10.1109/TIM.2014.2381354

[58] Jiacheng Shang, Si Chen, and Jie Wu. 2018. Defending Against Voice Spoofing: A Robust Software-based Liveness Detection System. In *2018 IEEE 15th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*. IEEE, 28–36.

[59] Jiacheng Shang, Si Chen, and Jie Wut. 2018. SRVoice: A Robust Sparse Representation-based Liveness Detection System. In *2018 IEEE 24th International Conference on Parallel and Distributed Systems (ICPADS)*. IEEE, 291–298.

[60] Robert V Shannon, Fan-Gang Zeng, Vivek Kamath, John Wygonski, and Michael Ekelid. 1995. Speech recognition with primarily temporal cues. *Science* 270, 5234 (1995), 303–304.

[61] Jan Silovskỳ and Jan Nouza. 2006. Speech, speaker and speaker's gender identification in automatically processed broadcast stream. *Radioengineering* (2006).

[62] J Singh, B Ginsburg, S Rao, and K Ramasubramanian. 2017. AWR1642 mmWave sensor: 76–81-GHz radar-on-chip for short-range radar applications. *Texas Instruments* (2017), 1–7.

[63] Craig S. Smith. [n.d.]. Alexa and Siri Can Hear This Hidden Command. You Can't. (Published 2018). http://www.nytimes.com/2018/05/10/technology/alexa-siri-hidden-command-audio-attacks.html

[64] synopsys [n.d.]. High-Performance DSP and Control Processing for Complex 5G Requirements. https://www.synopsys.com/designware-ip/technical-bulletin/high-performance-dsp-for-5g-dwtb-q418.html

[65] Guochao Wang, Jose-Maria Munoz-Ferreras, Changzhan Gu, Changzhi Li, and Roberto Gómez-García. 2014. Application of linear-frequency-modulated continuous-wave (LFMCW) radars for tracking of vital signs. *IEEE transactions on microwave theory and techniques* 62, 6 (2014), 1387–1399.

[66] Jianglin Wang. 2013. Physiologically-motivated feature extraction methods for speaker recognition. (2013).

[67] Qian Wang, Xiu Lin, Man Zhou, Yanjiao Chen, Cong Wang, Qi Li, and Xiangyang Luo. 2019. VoicePop: A pop noise based anti-spoofing system for voice authentication on smartphones. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE, 2062–2070.

[68] Teng Wei, Shu Wang, Anfu Zhou, and Xinyu Zhang. 2015. Acoustic eavesdropping through wireless vibrometry. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*. ACM, 130–141.

[69] Chenhan Xu, Zhengxiong Li, Hanbin Zhang, Aditya Singh Rathore, Huining Li, Chen Song, Kun Wang, and Wenyao Xu. 2019. WaveEar: Exploring a mmWave-based Noise-resistant Speech Sensing for Voice-User Interface. In *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 14–26.

[70] Y. Xu, S. Wu, C. Chen, J. Chen, and G. Fang. 2012. A Novel Method for Automatic Detection of Trapped Victims by Ultrawideband Radar. *IEEE Transactions on Geoscience and Remote Sensing* 50, 8 (Aug 2012), 3132–3142. https://doi.org/10.1109/TGRS.2011.2178248

[71] Chen Yan, Yan Long, Xiaoyu Ji, and Wenyuan Xu. 2019. The Catcher in the Field: A Fieldprint based Spoofing Detection for Text-Independent Speaker Verification. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. 1215–1229.

[72] Zhicheng Yang, Parth H Pathak, Yunze Zeng, Xixi Liran, and Prasant Mohapatra. 2016. Monitoring vital signs using millimeter wave. In *Proceedings of the 17th ACM International Symposium on Mobile Ad Hoc Networking and Computing*. ACM, 211–220.

[73] Xuejing Yuan, Yuxuan Chen, Yue Zhao, Yunhui Long, Xiaokang Liu, Kai Chen, Shengzhi Zhang, Heqing Huang, XiaoFeng Wang, and Carl A Gunter. 2018. Commandersong: A systematic approach for practical adversarial voice recognition. In *27th {USENIX} Security Symposium ({USENIX} Security 18)*. 49–64.

[74] Maxim Zhadobov, Nacer Chahat, Ronan Sauleau, Catherine Le Quement, and Yves Le Drean. 2011. Millimeter-wave interactions with the human body: state of knowledge and recent advances. *International Journal of Microwave and Wireless Technologies* 3, 2 (2011), 237âĂŞ247. https://doi.org/10.1017/S1759078711000122

[75] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. 2017. Dolphinattack: Inaudible voice commands. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. 103–117.

[76] Linghan Zhang, Sheng Tan, and Jie Yang. 2017. Hearing your voice is not enough: An articulatory gesture based liveness detection for voice authentication. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 57–71.

[77] Linghan Zhang, Sheng Tan, Jie Yang, and Yingying Chen. 2016. Voicelive: A phoneme localization based liveness detection for voice authentication on smartphones. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 1080–1091.

[78] Mingmin Zhao, Fadel Adib, and Dina Katabi. 2016. Emotion recognition using wireless signals. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*. ACM, 95–108.

[79] Mingmin Zhao, Shichao Yue, Dina Katabi, Tommi S Jaakkola, and Matt T Bianchi. 2017. Learning sleep stages from radio signals: A conditional adversarial architecture. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 4100–4109.

[80] Bing Zhou, Jay Lohokare, Ruipeng Gao, and Fan Ye. 2018. EchoPrint: Two-factor Authentication using Acoustics and Vision on Smartphones. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*. ACM, 321–336.